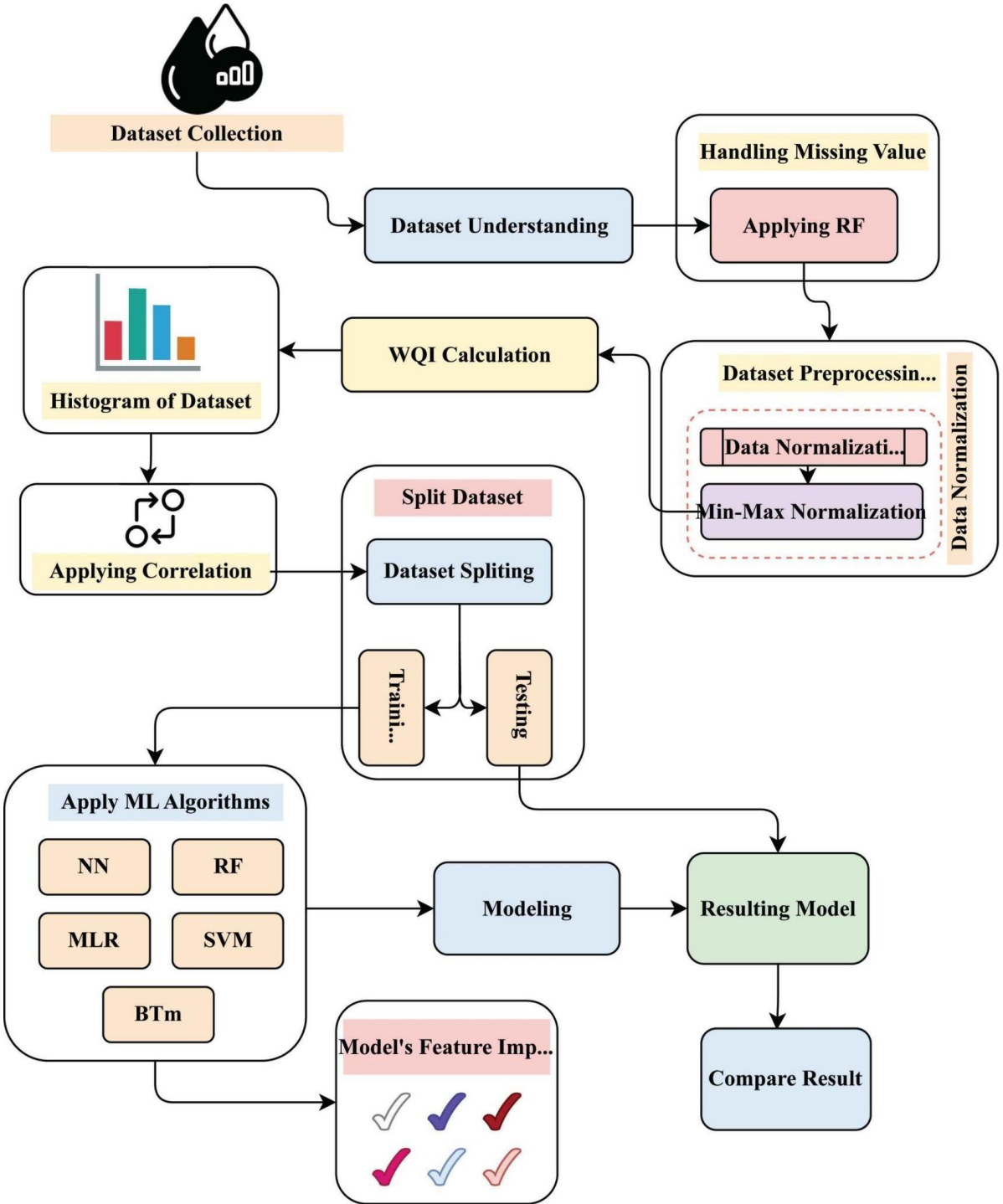
DATA FLOW DIAGRAM

Date	20.10.2022
Team ID	PNT2022TMID46440
Project Name	Efficient water quality analysis and prediction using machine learning
Maximum Marks	2 marks

To Undertake our Analysis, we initially have collected datasets from the Kaggle website, where we have used Indian water quality data. After which the dataset is properly analysed . Then, in order to counteract our vast number of datasets, we use many algorithms and some



procedures such as:

- **Handling Missing Data** :Random Forest, naive Bayes and k-NearestNeighbors to handle any missing values from such sets of data. We initially focus on data Normalization and then Min-Max Normalization, which is a crucial procedure in data analysis, to improve data quality for all of our datasets.

- **Calculate WQI** :Water Quality Index should be correct to assesswater quality in order to attain our purpose.

- **Visualisation :**We provide a Histogram of the dataset for better representation, where we can view the distribution of our entiredataset. Then we apply correlation to measure the changing capability of two variables at a constant rate.

- **Training and Test set Splitting :** Following that, the dataset isdistributed into two portions: training and testing.

- **Machine Learning Algorithms :** We have implemented five machine learning algorithms to train the dataset: Neural Network, Random Forest, Multinomial Logistic Regression, Support Vector Machine, and Bagged Tree Model, with six distinct colour signals indicating the relevance of each feature. Following the use of thosetechniques, we concentrate on modelling and obtain results from our intended model, which is also based on our dataset testing part.

- **Comparision:** Finally, we compare all the outcomes by using the accuracy of our suggested models and perhaps some current approaches. As a result, this methodology ensures the validity andtrustworthiness of our whole study.